DOCUMENT RESUME

ED 442 280 FL 026 284

AUTHOR Cheng, Liying
TITLE Washback or Backwash: A Review of the Impact of Testing on Teaching and Learning.
PUB DATE 2000-00-00
NOTE 34p.
PUB TYPE Opinion Papers (120)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Curriculum Based Assessment; Elementary Secondary Education; *Evaluation Research; Language Tests; *Performance Based Assessment; Second Language Instruction; Second Language Learning; *Testing
IDENTIFIERS *Teaching to the Test

ABSTRACT

Washback or backwash, also known as measurement-driven instruction, is a common term in applied linguistics referring to the influence of testing on teaching and learning, which is a prevailing phenomena in education. It is a truism that "what is assessed becomes what is valued, which becomes what is taught." This paper aims to share the discussion of this education phenomenon from different perspectives both in the area of general education and in language education. It discusses the historical origins of washback; the definition and scope of washback; and the function and mechanism of washback, and efforts, both recent and not, to mitigate its negative effects. It is concluded that the ultimate reason for the persistence and widespread nature of this problem is the existence of high-stakes testing. Few educators would dispute the claim that these sorts of high-stakes tests markedly influence the nature of instructional programs. Whether they are concerned about their own self-esteem or their students' well-being, teachers clearly want students to perform well on such tests. Accordingly, teachers tend to focus a significant portion of their instructional activities on the knowledge and skills assessed by such tests. (KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Liying Cheng

Washback or backwash:

A review of the impact of testing on teaching and learning

Liying Cheng, Ph.D.

Queen's University

ED 442 280

Summary

Washback or backwash, a common term used in applied linguistics, refers to the influence of testing on teaching and learning, which is a prevailing phenomena in education - 'what is assessed becomes what is valued, which becomes what is taught'. (McEwan, 1995a: 42). This review aims to share the discussion of this education phenomenon from different perspectives both in the area of general education and in language education. It discusses the origin of washback; the definition and scope of washback; and the function and mechanism of washback being exhorted for top-down educational policy, reform and accountability in many parts of the world.

Key words: washback, public examination, teaching and learning, accountability

The origin of washback

Washback (Alderson and Wall, 1993) ***or backwash*** (Biggs, 1995, 1996), refers to as the influence of testing on teaching and learning, which is rooted in the notion that tests *should* and *could* drive teaching and hence learning (referred as ***measurement-driven instruction*** by Popham, 1983,1987). In order to achieve the goal, a 'match' or an overlap between the content and format of the test and the content and format of the curriculum (or curriculum surrogate such as the textbook) is encouraged (***curriculum alignment*** by

BEST COPY AVAILABLE

Shepard, 1990;1991;1992;1993). The closer the fit or match, the greater the potential improvement on the test. However, the idea of alignment – matching the test and curriculum – has been declared as 'unethical' (Haladyna et al, 1991:4; Widen et al, 1997 for example). Such alignment is particularly evident in many places of the world, e.g. the Hong Kong context (especially in terms of the textbooks) (see Cheng, 1997a, 1997b), where new examinations or revised examinations are introduced into the education system to improve teaching and learning (***systemic validity*** and ***consequential validity*** by Fredericksen and Collins, 1989; Messick, 1989, 1992, 1994, 1996). Bachman and Palmer (1996) and Baker (1991) refer to the phenomena as ***test impact***. These above terms and possible other terms all refer to different aspects of the same phenomenon - the influence of testing on teaching and learning.

The study of washback has indeed been derived from recent developments in language testing and measurement-driven reform on instruction in general education. Research in language testing has centred around whether and how we assess the specific characteristics of a given group of test-takers and whether and how we can incorporate such information into the way we design language tests. Perhaps the single most important theoretical development in language testing since the 1980's was the realisation that a language test score represents a complexity of multiple influences. Language test scores cannot be interpreted simplistically as an indicator of the particular language ability we want to measure. They are also affected by the characteristics and content of the test tasks, the characteristics of the test taker, the strategies the test taker employs in attempting to complete the test task, and the inferences we wish to draw from them (referred to as consequential validity by Messick, 1989). What makes the

interpretation of test scores particularly difficult is that these factors undoubtedly interact with each other.

Following the above discussion, Alderson (1986) pointed out '*washback*' as an additional area to which language testing needed to turn its attention to in the years to come. Alderson (1986:104) discussed the 'potentially powerful influence offsets', and argued for innovations in the language curriculum through innovations in language testing (see also Wall, 1996). Davies (1985) asked whether tests necessarily would follow the curriculum, and suggested that perhaps tests ought to *lead* and *influence* curriculum. Morrow (1986:6) further used the term '*washback validity*' to describe the quality of the relationship between testing and teaching and learning. He claimed that '... in essence an examination of washback validity would take testing researchers into the classroom in order to observe the effect of their tests in action.' This has important implications for test validity.

Messick (1989, 1992, 1994, 1996) has placed the washback effect within a broader concept of construct validity (*consequential validity*) (see also Cronbach, 1988). Messick claimed that construct validity encompasses aspects of test use, the impact of tests on test takers and teachers, the interpretation of scores by decision makers, and the misuses, abuses, and unintended uses of tests. Washback is an inherent quality of any kind of assessment, especially when people's futures are affected by the examination results, regardless of the quality of the examination (Eckstein and Noah, 1992, 1993a, 1993b).

Examinations have been long used as means of control. They have been with us for a long time, at least a thousand years or more, if the use made of them in Imperial China to

select the highest officials of the land is included (Arnove, Altback and Kelly, 1992; Lai, 1970; Hu, 1984). Those used was probably the first Civil Service Examination ever developed by our human race. To avoid corruption, all essays in the Imperial Examination were marked anonymously, and the Emperor personally supervised the final stage. Although the goal of the examination was to select civil servants, its washback effect was to establish and control an educational program, as prospective mandarins set out to prepare themselves for the examination (Spolsky, 1994, 1995).

Even in modern times, the use of examinations to select for education and employment dates back at least 300 years. Examinations were seen as ways to encourage the development of talent, to upgrade the performance of schools and colleges, and to counter, to some degree, nepotism, favouritism, and even outright corruption in the allocation of scarce opportunities (Eckstein and Noah, 1992; Bray and Steward, 1998). If the initial spread of examinations can be traced to such motives, the very same rationales appear to be as powerful as ever today. Examinations are subject to much criticism. However, in spite of all the criticism levelled at them, examinations continue to occupy a leading place in the educational arrangement of most countries these days (Baker, 1991; Calder, 1990, 1997; Cannell, 1987; Cheng, 1997a, 1997b; Heyneman, 1987; Heyneman and Ransom, 1990; Kellaghan and Greaney, 1992; Li, 1990; Macintosh, 1986; Runte, 1998; Shohamy, 1993; Shohamy, et al 1996; Widen et al, 1997).

Policy-makers in central agencies especailly, aware of the power of tests, use them to manipulate educational systems, to control curricula and to impose new textbooks and new teaching methods. In those centralised countries, tests are viewed as the primary tools through which changes in the educational system can be introduced without having

to change other educational components such as teacher training or curricula, which further demonstrates the high-stakes role of examinations in education. Furthermore, Shohamy et al (1996:299) stated that 'the power and authority of tests enable policy-makers to use them as effective tools for controlling educational systems and prescribing the behaviour of those who are affected by their results - administrators, teachers and students. School-wide exams are used by principals and administrators to enforce learning, which in classrooms, tests and quizzes are used by teachers to impose discipline and to motivate learning' (Stiggins and Faires-Conklin, 1992).

Consequently, testing has become 'the darling of policy makers' across the country under the educational system in the USA (Madaus 1985a, 1985b). Similar statements could have been made at various times during the past century and a half, most notably during periods when schools were under attack and reformers sought to demonstrate the need for change (Linn, 1992). In Canada, a consortium of provincial ministers of education recently instituted a project of national achievement testing in the areas of reading, language arts, and science (Council of Ministers, 1994). Several provinces such as British Columbia, Alberta, Saskatchewan, Quebec, and Newfoundland require students to pass centrally set school-leaving examinations as a condition for school graduation (see Anderson et al 1990, Runte, 1998; Widen et al, 1997). Petric (1987:175) emphasises that 'it would not be too much of an exaggeration to say that evaluation and testing have become *the* engine for implementing educational policy'.

Popham (1987) outlined the traditional notion of measurement-driven instruction to illustrate the relationship between instruction and assessment: assessment directs teachers' attention to the content of test items, acting as powerful 'curricular magnets'. In

high-stakes environments, in which the results of mandated tests trigger rewards, sanctions, or public scrutiny and loss of professional status, teachers will be motivated to pursue the objectives that the test embodies. Given the important decisions attached to examinations, it is natural that they have always been used as instruments and targets of control in school systems (Eckstein and Noah, 1993a, 1993b; Smith et al. 1990, Wesdrop, 1982). Their relationship with the curriculum, with teacher teaching and student learning, and to individual life chances are of vital importance in most societies.

The definition and scope of washback

Washback is a term commonly used in applied linguistics, yet it is rarely found in dictionaries. However, the word 'backwash' can be found in certain dictionaries and is defined as 'the unwelcome repercussions of some social action' by the New Webster's Comprehensive Dictionary of the English Language, and 'unpleasant after-effects of an event or situation' by Collin Cobuild Dictionary of English Language.

Washback is defined as '...[the impact of a test on teaching] and ... tests can be powerful determiners, both *positively and negatively*, of what happens in classrooms' (Wall and Alderson 1993:41). It refers to the extent to which the test influences language teachers and learners to do things 'they would not necessarily *otherwise* do because of the test' (Alderson and Wall 1993:117). Messick (1996:241) emphasises that 'washback, a concept prominent in applied linguistics, refers to the extent to which the introduction and the use of a test influences language teachers and learners to do things they would not otherwise do that promote or inhibit language learning.' He continues to comment that 'some proponents have even maintained that a test's validity should be appraised by the

degree to which it manifests positive or negative washback, a notion akin to the proposal of 'system validity' (Frederiksen and Collins, 1989) in the educational measurement literature. Shohamy notes (1992:513) that 'this phenomenon is the result of the strong authority of external testing and the major impact it has on the lives of test takers'. Pearson (1988:98) points out that 'public examinations influence the attitudes, behaviours, and motivation of teachers, learners and parents, and, because examinations often come at the end of a course, this influence is seen working in a backward direction - hence the term 'washback'. He further emphasises that the direction in which washback actually works must be *forwards* in time.

Biggs (1995:12) uses the term 'backwash' to refer to the fact that testing drives not only the curriculum, but teaching methods and students' approaches to learning (Crooks, 1988; Frederiksen, 1984; Frederiksen and Collins, 1989). However, quoting definitions of the term 'backwash' from the Collin Cobuild Dictionary of English Language, Spolsky (1994:55) commented that 'backwash is better applied only to accidental side-effects of examinations, and not to those effects intended when the first purpose of the examination is control of the curriculum'. Cheng (1997a, 1997b, 1999) prefers the term of 'washback' in an empirical study of an intended public examination change on classroom teaching. She defines the phenomenon as 'an intended direction and function of curriculum change by means of a change of public examinations on aspects of teaching and learning'. However, it should be pointed out that when public examinations are used as a vehicle for any intended curriculum change, unintended and accidental side-effects also happen, as successful curriculum change and development is a highly complex matter. A study into the phenomenon needs to chart the on-going process of investigating public examinations

in an given education context by exploring 'where', the school or university contexts, 'when', the time duration of using such assessment practices, 'why', the rationale and 'how', the different approaches used by different participants within the context. Consequently, both intended and unintended effect could happen. According to Alderson and Wall (1993:115), the notion that testing influences teaching is referred to as 'backwash' in general educational circles, but it has come to be known as 'washback' among British applied linguists, though they see no reason, semantic or pragmatic, for using either term. I have kept the term 'washback' or 'backwash' as it appears in each study.

Messick (1996:241) further discusses that 'for optimal positive washback there should be little, if any difference between activities involved in learning the language and activities involved in preparing for the test'. However, 'such forms of evidence are only circumstantial with respect to test validity in that a poor test may be associated with positive effects and a good test with negative effects because of other things that are done or not done in the education system' (Messick, 1996:242). However, Alderson and Wall (1993:116) argue that 'washback, if it exists - which has yet to be established - is likely to be a complex phenomenon which cannot be related directly to a test's validity'. The washback effect should refer to the effect of the test itself on aspects of teaching and learning. Besides, other operating forces within the education context also contribute to or ensure the washback effect on teaching and learning, which proves to be true when we look at various washback studies (see Anderson et al, 1990; Cheng, 1998, 1999, Herman, 1992; Madaus, 1988; Smith, 1991, Wall et al, 1996, Watanabe, 1996; Widen et al, 1997).

Bailey (1996:259) summarises, after considering several definitions of washback, that

washback is generally defined as the influence of testing on teaching and learning: in which it is widely held to exist and to be important, but relatively little empirical research has been done to document its exact nature or the mechanisms by which it works. She commented further that 'there are also concerns about what constitutes both positive and negative washback, as well as about how to promote the former and inhibit the latter.'

- **Negative Washback**

Language tests and tests in general are often criticised for their negative influence on teaching - so-called 'negative washback' (see Alderson and Wall 1993:115). Vernon (1956:166) commented that teachers tended to ignore subjects and activities, which did not contribute directly to passing the exam, and claimed that examinations 'distort the curriculum'. Davies (1968a:125, 1968b), for example, indicates that all too often the washback effect has been bad: designed as testing devices, examinations have become teaching devices; work is directed to what are - in effect if not in fact - past examination papers and consequently becomes narrow and uninspired. Alderson and Wall (1993:5) refer to 'negative washback' as the negative or undesirable effect on teaching and learning of a particular and, by inference if not direct statement, 'poor' test. In this case, 'poor' usually means 'something that the teacher or learner does not wish to teach or learn.' The tests may well fail to reflect the learning principles and/or the course objectives to which they are supposedly related.

Fish (1988) discovered that teachers reacted negatively to pressure created by public displays of classroom scores, and also found that relatively inexperienced teachers felt greater anxiety and accountability pressure than did experienced teachers. Noble and

Smith (1991a:3) also pointed out that high-stakes testing affected teachers directly and negatively, and that 'teaching test-taking skills and drilling on multiple-choice worksheets is likely to boost the scores but unlikely to promote general understanding'(1991b:6). Smith (1991b:8) concluded from an extensive qualitative study of the role of external testing in elementary schools that 'testing programs substantially reduce the time available for instruction, narrow curricular offerings and modes of instruction, and potentially reduce the capacities of teachers to teach content and to use methods and materials that are incompatible with standardised testing formats'.

According to Anderson et al (1990) survey study in British Columbia, when investigating the impact of re-introducing final examinations at Grade12, teachers reported a narrowing to the topics the examination was most likely to include, and that student adopted more of a memorisation approach, with reduced emphasis on critical thinking. The Grade 12 examinations have affected students in lower grades through increased school-wide tests, increased emphasis on test-taking skills, and increased attention to subject matter associated with the examination. In another study (Widen et al, 1997), Grade 12 teachers believe that they have lost much of their discretion in curriculum decision making, and, therefore, much of their autonomy. When teachers are being circumscribed and controlled by the examinations, and students' whole only focus is on what will be tested, teaching is limited to the testable aspects of the discipline (see also Calder, 1990, 1997).

- **Positive Washback**

Some researchers, on the other hand, strongly believe that it is feasible and desirable to bring about beneficial changes in language teaching by changing examinations - so-called

'positive washback'. This term is directly related to 'measurement-driven instruction' in general education, and refers to tests that influence teaching and learning beneficially (see Alderson and Wall, 1993:115). In this sense, teachers and learners have a positive attitude toward the test and work willingly toward its objectives. Pearson (1988:107) argued that 'good tests will be more or less directly usable as teaching-learning activities. Similarly, good teaching-learning tasks will be more or less directly usable for testing purposes, even though practical or financial constraints limit the possibilities'. Considering the complexity of teaching and learning, such claim sounds ideal, but rather simplistic. Davies (1985) maintains the view that a good test should be 'an obedient servant of teaching; and this is especially true in the case of achievement testing'. He (1985:8) further argues that 'creative and innovative testing ... can, quite successfully, attract to itself a syllabus change or a new syllabus which effectively makes it into an achievement test.' In this case, the test no longer needs to be only an 'obedient servant': rather it can also be a 'leader'.

However, there are rather conflicting reactions toward whether there is positive or negative washback on teaching and learning. Wiseman (1961:159) argued that paid coaching classes, which were intended for preparing students for exams, were not a good use of the time, because students were practising exam techniques rather than language learning activities. However, Heyneman (1987:262) commented that many proponents of academic achievement testing view 'coachability' not as a drawback, but rather as a virtue. Pearson (1988:101) looks at the washback effect of a test from the point of view of its potential negative and positive influences on teaching. According to him, a test's washback effect will be negative if it fails to reflect the learning principles, and/or course

objectives to which they supposedly relate, and it will be positive if the effects are beneficial and 'encourage the whole range of desired changes'.

Alderson and Wall (1993:118-117), on the other hand, stress that the quality of the washback effect might be independent of the quality of a test. Any test, good or bad, can be said to have beneficial or detrimental washback. Whatever changes educators would like to bring about in teaching and learning by whatever assessment methods, it is worthwhile to investigate first the broad educational context in which an assessment is introduced since other forces exist within the society, education, and schools that might prevent washback from appearing (Alderson and Wall, 1993: 116).

Heyneman (1987:262) concluded that 'testing is a profession, but it is highly susceptible to political interference. To a large extent, the quality of tests relies on the ability of a test agency to pursue professional ends autonomously'. If the consequences of a particular test for teaching and learning are to be evaluated, the educational context in which the test takes place needs to be investigated. Whether the washback effect is positive or negative will largely depend on *how it works* and *within which educational contexts*.

The function and mechanism of washback

Traditionally, tests come at the end of the teaching and learning process for evaluative purposes. However, with the advent of high-stakes public examination system nowadays, the direction seems to be reversed. Testing usually comes first before the teaching and learning process. When examinations are commonly used as levers for change, new textbooks will be designed to match the purposes of a new test, and school administrative and management staff, teachers and students will work harder to achieve good scores on

the test. In addition, many more changes in teaching and learning can happen as a result of a particular new test. However, the consequences may be independent of the original intentions of the test designers.

Shohamy (1993:2) pointed out that 'the need to include aspects of test use in construct validation originates in the fact that testing is not an isolated event; rather, it is connected to a whole set of variables that interact in the educational process'. Moreover, Messick (1989) recommended a unified validity concept, in which he shows that when an assessment model is designed to make inferences about a certain construct, the inferences drawn from that assessment model should not only derive from test score interpretation but also from other variables in the social context (Bracey, 1989; Cooley 1991; Cronbach, 1988; Gardner, 1992; Gifford and O' Connor 1992; Linn, Baker and Dunbar, 1991; Messick, 1992). As early as 1975, Messick (1975:6) pointed out that 'researchers, other educators, and policy makers must work together to develop means of evaluating educational effectiveness that accurately represent a school or district's progress toward a broad range of important educational goals.' In this context, Linn (1992:29) stated that it is incumbent upon the measurement research community to make the case that the introduction of any new high-stakes examination system should include more provisions for paying greater attention to investigations of both the intended and unintended consequences of the system than has been typical of previous test-based reform efforts.

Exploring the mechanism of such an assessment function, Bailey (1996:262-264) cites Hughes' trichotomy (1993) to illustrate the complex mechanism by which washback works in actual teaching and learning context.

Table 1 The trichotomy of backwash model (Source: Hughes, 1993:2)

1. participants - students, classroom teachers, administrators, materials developers and publishers, whose perceptions and attitudes towards their work may be affected by a test
2. process - any actions taken by the participants which may contribute to the process of learning
3. product - what is learned and the quality of the learning

Hughes (1993:2) further notes:

> *The trichotomy ... allows us to construct a basic model of backwash. The nature of a test may first affect the perceptions and attitudes of the participants towards their teaching and learning tasks. These perceptions and attitudes in turn may affect what the participants do in carrying out their work (process), including practising the kind of items that are to be found in the test, which will affect the learning outcomes, the product of the work.*

While Hughes focused on participants, processes and products in his backwash model to illustrate the washback mechanism, Alderson and Wall (1993), in their Sri Lankan study, focused on micro aspects of teaching and learning that might be influenced by examinations. They come up with 15 hypotheses regarding washback (1993:120-21) to illustrate areas in teaching and learning that usually receive washback. In addition, Cheng (1997a, 1997b), through a large-scale quantitative and qualitative empirical study,

developed the notion of 'washback intensity' to refer to the degree of the washback effect in an area or a number of areas of teaching and learning affected by an examination. Each of the areas ought to be studied in the future studies in order to chart and understand the function and mechanism of washback - the participants, the process and the products - that might be brought about by the change of a major public examination.

- A test will influence teaching.
- A test will influence learning.
- A test will influence what teachers teach.
- A test will influence how teachers teach.
- A test will influence what learners learn.
- A test will influence how learners learn.
- A test will influence the rate and sequence of teaching.
- A test will influence the rate and sequence of learning.
- A test will influence the degree and depth of teaching.
- A test will influence the degree and depth of learning.
- A test will influence attitude to the content, method, etc. of teaching and learning.
- Tests that have important consequences will have washback; and conversely
- Tests that do not have important consequences will have no washback.
- Tests will have washback on all learners and teachers.
- Tests will have washback effects for some learners and some teachers, but not for

others.

Alderson and Wall concluded that further research on washback is needed, and that such research must entail 'increasing specification of the Washback Hypothesis' (1993:127). They called on researchers to take account of findings in the research literature in at least two areas: 1) motivation and performance, 2) innovation and change in the educational settings.

Wall (1996:334), following up the above study, stressed the difficulties in finding explanations of how tests exert influence on teaching. She took from the innovation literature and added into her research areas (Wall et al 1996) to explain the complexity of the phenomenon.

- The writing of detailed baseline studies to identify important characteristics in the target system and the environment, including an analysis of current testing practices (Shohamy et al, 1996), current teaching practices, resources (Bailey, 1996; Stevenson and Riewe, 1981), and attitudes of key stakeholders (Bailey, 1996; Hughes, 1993).

- The formation of management teams representing all important interest groups: teachers, teachers trainers, university specialists and ministry officials (Cheng 1997a, 1997b).

Fullan (1991, 1993), also in the context of innovation, discussed changes in schools and came up with two major themes:

- Innovation should be seen as a process rather than as an event (p. 47)

- All the participants who are affected by an innovation have to find their own 'meaning' for the change (p.30).

He explained that the 'subjective reality', which teachers experience, would always contrast with the 'objective reality' that the proponents of change has originally imagined. Teachers work on their own, with little reference to experts or consultation with colleagues. They are forced to make on-the-spot decisions, with little time to reflect on better solutions. They are pressured to accomplish a great deal, but are given far too little time to achieve their goals. When, on top of this, they are expected to carry forward an innovation that someone else has come up with, their lives can become very difficult indeed (see also Huberman and Miles, 1984). Besides, it is also found that there tends to be discrepancies between the intention of any innovation or curriculum change and the understanding of teachers (Andrews and Fullilove, 1994; Markee, 1997).

Andrews (1994a, 1994b) highlighted the reality of the relationship between washback and curriculum innovation and summarised three choices by which educators might deal with washback: fight it, ignore it, or use it (see also Heyneman, 1987:260). By fighting it, Heyneman refers to the effort to replace examinations by other sorts of selection criteria, on the grounds that examinations have encouraged rote memorisation at the expense of more desirable educational practices. Andrews (1994b: 51-52) used the metaphor of the ostrich for those who ignore it. Those who are involved with mainstream activities, such as syllabus design, material writing and teacher training view testers as a 'special breed' using an arcane terminology. Tests and exams have been seen as an occasional necessary evil, a dose of unpleasant medicine, the taste of which should be washed away as quickly as possible. By using washback, the purpose is to promote pedagogical ends, which is not

a new idea in education (see also Andrews and Fullilove 1993, 1994; Blenkin, Edwards and Kelly, 1992; Brooke and Oxenham, 1984; Pearson, 1988; Somerset, 1983; Swain, 1984).

The function of assessment is generally believed to leverage educational change, which has often led to top-down educational reform strategies by employing 'better' kinds of assessment practices (Noble & Smith 1994a). Assessment practices are currently undergoing a major paradigm shift in many parts of the world. It can be described as a reaction to the perceived shortcomings of the prevailing paradigm with its emphasis on standardised testing (Biggs, 1992, 1996; Genesee, 1994). Alternative assessment methods have thus emerged as a systematic attempt to measure a learner's ability to use previously acquired knowledge in solving novel problems or completing specific tasks. Such assessment has been initiated to reflect a trend towards using assessment to reform curriculum and improve instruction at the school level (also refereed to as intended washback effect) (Noble and Smith, 1994a, 1994b; Popham, 1983 , 1987; Linn, 1983, 1992).

According to Noble and Smith (1994b:1), 'the most pervasive tool of top-down policy reform is to mandate assessment that can serve as both guideposts and accountability.' (see also Baker, 1989; Herman, 1989, 1992; McEwen, 1995a, 1995b, Resnick and Resnick, 1992; Resnick, 1989). They also point out that the goal of current measurement-driven reform in assessment is to build a better test that will drive schools toward more ambitious goals and reform them toward a curriculum and pedagogy geared more toward thinking and less toward rote memory and isolated skills - the shift from behaviourism to cognitive-constructivism in teaching and learning beliefs.

Beliefs about testing tend to follow beliefs about teaching and learning (Glaser and Bassok, 1989; Glaser and Silver, 1994). According to the more recent psychological and pedagogical view of learning, labelled cognitive-constructivism, effective instruction must mesh with how students think. The direct instruction model under the influence of behaviourism - tell-show-do approach - does not match how students learn, nor does it take into account pupil intention, interest, and choice. Teaching that fits the cognitive-constructivist view of learning is likely to be holistic, integrated, project-oriented, long-term, discovery-based, and social, so should testing be. Thus cognitive-constructivists see performance assessment[1] as parallel to the above belief of how pupils learn and how they should be taught. Performance-based assessment can be designed to be so closely linked to the goals of instruction as to be almost indistinguishable from them. Rather than being a negative consequence, as it is now with some high-stakes uses of existing standardised tests, 'teaching to these proposed performance assessments, accepted by scholars as inevitable and by teachers as necessary, becomes a virtue, according to this line of thinking' (Noble and Smith, 1994b:7; see also Aschbacher, 1990; Aschbacher, Backer and Herman, 1988; Baker et al, 1992; Wiggins, 1989a, 1989b, 1993). The rational also lay in the belief that measurement-driven instruction was initiated due to public discontent with the quality of schooling (Popham, Rankin, Standifer and Williams, 1985:629).

However, such a reform strategy was counter-argued by Andrews (1994a & b) as a 'blunt instrument' for bringing about changes in teaching and learning since the actual teaching

[1] Performance assessment based on the constructivist model of learning is defined by Gipps as (1994:99) 'a systematic attempt to measure a learner's ability to use previously acquired knowledge in solving novel problems or completing specific tasks. In performance assessment, real life or simulated assessment exercises are used to elicit original responses, which are directly observed and rated by a qualified judge'.

and learning situation is clearly far more complex as discussed above than proponents of alternative assessment suggest (see also empirical washback studies by Alderson and Wall 1993, Cheng, 1997b, 1999, Wall, 1996). Each different educational context (school environment, messages from administration, expectations of other teachers, and students) plays a key role in facilitating or detracting from the possibility of change. Therefore, such a strategy seems rather simplistic. Besides, Noble and Smith (1994a:1-2), in their study of the impact of the Arizona Student Assessment Program, revealed both the ambiguities of the policy-making process and the dysfunctional side effects that evolved from the policy's disparities, though the legislative passage of the testing mandate obviously demonstrated Arizona's commitment to top-down reform and its belief that assessment can leverage educational change. The relationship between testing and teaching and learning is obviously more complicated than just the design of a 'good' assessment type . There is more underlying interplay and intertwining within each specific educational context where the assessment takes place.

Furthermore, Madaus (1988:85) emphasises:

> *The tests can become the ferocious master of the educational process, not the compliant servant they should be. Measurement-driven instruction invariably leads to cramming; narrows the curriculum; concentrates attention on those skills most amenable to testing; constrains the creativity and spontaneity of teachers and students; and finally demeans the professional judgement of teachers. (1988:85)*

According to Madaus (1988), a high-stakes test can lever the development of new

curricular materials, which can be a positive aspect. However, even if new materials are produced as a result of a new examination, they might not be moulded according to the innovators' view of what is desirable in terms of teaching, but might rather according to the publishers' view of what will sell (see Andrews, 1994b; Cheng, 1997b), which is in fact the situation within the Hong Kong education context.

In despite, measurement-driven instruction will occur when a high-stakes test of educational achievement influences the instructional program that prepares students for the test (Popham 1987:680) since important contingencies are associated with the students' performance in such a situation.

> *Few educators would dispute the claim that these sorts of high-stakes tests markedly influence the nature of instructional programs. Whether they are concerned about their own self-esteem or their students' well-being, teachers clearly want students to perform well on such tests. Accordingly, teachers tend to focus a significant portion of their instructional activities on the knowledge and skills assessed by such tests. (1987:680)*

In the end, the change is in the teachers' hands. As English (1992) pointed out, when the classroom door is shut and nobody else is around, the classroom teacher can then select and teach almost any curriculum he or she decides is appropriate irrespective of the various reforms, innovations and public examinations.

References:

ALDERSON, J. C. (1986). 'Innovations in language testing.' In: PORTAL M.(Ed.) *Innovations in Language Testing: Proceedings of the IUS/NFER Conference*. Windsor: NFER-Nelson.

ALDERSON, J.C. and WALL, D. (1993). 'Does washback exist?', *Applied Linguistics*, **14**, 2, 115-129.

ANDERSON, J. O., MUIR, W., BATESON, D. J., BLACKMORE, D. and ROGERS, W. T. (1990). *The impact of provincial examinations on education in British Columbia: General report*. Report to the British Columbia: British Columbia Ministry of Education. Victoria: British Columbia Ministry of Education.

ANDREWS, S. and FULLILOVE, J. (1994). 'Assessing spoken English in public examinations - Why and how?' In: BOYLE, J. and FALVEY, P. (Eds.) *English Language Testing in Hong Kong*. Hong Kong: The Chinese University Press.

ANDREWS, S. and FULLILOVE, J. (1993). Backwash and the Use of English Oral Speculations on the impact of a new examination upon sixth form English language testing in Hong Kong. *New Horizons*, **34**,46-52.

ANDREWS, S. (1994a). The washback effect of examinations - Its impact upon curriculum innovation in English language teaching. *Curriculum Forum*, **4**, 1, 44-58.

ANDREWS, S. (1994b). Washback or washout? The relationship between examination reform and curriculum innovation. In NUNAN, D. BERRY, V. and BERRY, R.(Eds.). *Bringing about change in language education*. Hong Kong: Dept. of Curriculum Studies, University of Hong Kong, pp. 67-81.

ARNOVE, R. F., ALTBACK, P. G., and KELLY, G. P. (Eds.). (1992). *Emergent issues in education: Comparative perspectives*. Albany, NY: State University of New York Press.

ASCHBACHER, P. E., BAKER, E. L. and HERMAN, J. L. (Eds.). (1988). *Improving large-scale assessment*. Resource Paper No. 9. Los Angeles, CA: CRESST, UCLA Graduate School of Education.

ASCHBACHER, P. R. (1990, November). Monitoring the impact of testing and evaluation innovations projects: State activities and interest concerning performance-based assessment. Los Angeles, CA: CRESST.

BACHMAN, L. F. and PALMER, A. S. (1996). *Language Testing in Practice*. Oxford: Oxford University Press.

BAILEY, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, **13**, 3, 257-279.

BAKER, E. L. (1991). *Alternative assessment and national policy*. Paper presented at the National Research Symposium on Limited English Proficient Students' Issues: Focus on Evaluation and Measurement, Washington, DC.

BAKER, E. L. (1989). *Can we fairly measure the quality of education?* (Centre for Study of Evaluation Technical Report 290). Los Angeles, CA: Centre for Study of Evaluation (CSE).

BAKER, E., ASCHBACHER, P., NIEMI, D. and SATO, E. (1992). *Performance assessment models: Assessing content area explanations*. Los Angeles: University of California, CRESST.

BIGGS, J. B. (1992). The psychology of assessment and the Hong Kong scene. *Bulletin of the Hong Kong Psychological Society*, **27/28**, 1-21.

BIGGS, J. B. (1995). Assumptions underlying new approaches to educational assessment. *Curriculum Forum*, **4**, 2, 1-22.

BIGGS, J. B. (Ed.) (1996). *Testing: To Educate or to Select? Education in Hong Kong at the Cross-roads*. Hong Kong: Hong Kong Educational Publishing Co.

BLENKIN, G. M., EDWARDS, G. and KELLY, A.V. (1992). *Change and the curriculum*. London: P. Chapman.

BRACEY, G.W. (1989). The $150 million redundancy. *Phi Delta Kappa*, **70**, 9, 698-702.

BRAY, M. and STEWARD L. (Eds.) (1998). Examination Systems in Small States: comparative perspectives on policies, models and operations. London: Commonwealth Secretariat.

BROOKE, N. and OXENHAM, J. (1984). The influence of certification and selection on teaching and learning. In: Oxenham, J.(Ed.). *Education versus qualifications?* London: Allen and Unwin, pp. 147-75.

CALDER, P. (1990). *Impact of diploma examinations on the teaching-learning process*. Admonition: Alberta Teacher Association.

CALDER, P. (1997). *Impact of Alberta achievement tests on the teaching-learning process*. Admonition: Alberta Teacher Association.

CANNELL, J. J. (1987). Nationally-normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement: Issues and Practice*, **7**, 4, 12-15.

CHENG, L. (1997a). 'How does washback influence teaching? Implications for Hong Kong', *Language and Education*, **11**, 1, 38-54.

CHENG, L. (1997b). The washback effect of public examination change on classroom teaching: An impact study of the 1996 Hong Kong Certificate of Education in English on the classroom teaching of English in Hong Kong secondary schools. Unpublished Doctoral Dissertation. The University of Hong Kong: Hong Kong.

CHENG, L. (1998). 'Impact of a public English examination change on students' perceptions and attitudes toward their English learning', *Studies in Educational Evaluation*, **24**, 3, 279-301.

CHENG, L. (1999). 'Changing assessment: Washback on teacher perspectives and actions', *Teaching and Teacher Education* **15**, 3, 253-271.

COOLEY, W. W. (1991). State-wide student assessment. *Educational Measurement: Issues and Practice*, **10**, 3-6.

COUNCIL OF MINISTERS OF EDUCATION, CANADA. (1994). School Achievement indicators program. Toronto: Author.

CRONBACH, L. J. (1988). Five perspectives on the validity argument. In: WAINER H. and BRAUN, H. I. (Eds.) *Test Validity*. Hilldale, NJ: Erlbaum, pp.3-17.

CROOKS, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, **58**, 4, 438-481.

DAVIES, A. (1985). Follow my leader: Is that what language tests do? In: LEE Y.P. and [others] (Eds.) *New Directions in Language Testing*. Oxford: Pergamon Press, pp-12.

DAVIES, A. (Ed.). (1968). *Language testing symposium: A psycholinguistic approach*. London: Oxford University Press.

ECKSTEIN, M. A. and NOAH, H. J. (1993a). *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven: Yale University Press.

ECKSTEIN, M. A. and NOAH, H. J. (1993b). The politics of examinations: Issues and conflicts. In: ECKSTEIN M. A. and NOAH, H. J. (Eds.) *Secondary School Examinations: International Perspectives on Policies and Practice*. New Haven: Yale University Press, pp. 191-216.

ECKSTEIN, M. A. and NOAH, H. J. (Eds.) (1992). *Examinations: Comparative and International Studies*. Oxford: Pergamon Press.

FISH, J. (1988). *Responses to mandated standardised testing*. Unpublished doctoral dissertation, University of California, Los Angeles.

FREDERIKSEN, J.R. and COLLINS, A. (1989). A system approach to educational testing. *Educational Researcher*, **18**, 9, 27-32.

FREDERIKSEN, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychology*, **39**, 193-202.

FULLAN M. G. (1993). *Change forces: Probing the depth of educational reform*. London: Falmer Press.

FULLAN, M. G. with STIEGELBAUER, S. (1991). *The New Meaning of Educational Change*. (2nd edn). London: Cassell.

GARDNER, H. (1992). Assessment in context: The alternative to standardised testing. In: Gifford, B. R., and O'Connor, M. C. (Eds.). *Changing assessments: Alternative views of aptitude, achievement and instruction*. London: Kluwer Academic Publishers, pp. 77-119.

GENESEE, F. (1994). Assessment alternatives. *TESOL Matters* **3.**

GERGEN, K. J. (1985). The social constructionist movement in modern psychology. *American Psychologist*, **39**, 2, 93-104.

GIFFORD, B. R. and O'CONNOR, M. C. (Eds.). (1992). *Changing assessments: Alternative views of aptitude, achievement and instruction*. London: Kluwer Academic Publishers.

GIPPS, C. V. (1994). *Beyond testing: Toward a theory of educational assessment*. London: The Falmer Press.

GLASER, R. and BASSOK, M. (1989). Learning theory and the study of instruction. *Annual Review of Psychology*, **40**, 631-666.

GLASER, R. and SILVER, E. (1994). *Assessment, testing, and instruction: Retrospect and prospect* (CSE Technical Report 379). University of Pittsburgh: CRESST/ Learning Research and Development Centre.

HALADYNA, T. M., NOLEN, S. B. and HAAS, N. S. (1991). Raising standardised achievement test scores and the origins of test score pollution. *Educational Research, 20*(5), 2-7.

HERMAN, J. L. (1989). *Priorities of educational testing and evaluation: The testimony of the CRESST National Faculty* (CSE Technical Report 304). Los Angeles, CA: CRESST.

HERMAN, J. L. (1992). *Accountability and alternative assessment: Research and Development issues* (CSE Technical Report 384). Los Angeles, CA: CRESST.

HERMAN, J. L. and GOLAN, S. (1991). *Effects of standardised testing on teachers and learning - Another look* (CSE Technical Report 334). Los Angles, CA: CRESST.

HEYNEMAN, S. P. and RANSOM, A. W. (1990). Using examinations and testing to improve educational quality. *Educational Policy*, **4**, 3, 177-192.

HEYNEMAN, S. P. (1987). Use of examinations in developing countries: Selection, research, and education sector management. *International Journal of Education Development*, **7**, 4, 251-263.

HU, C. T. (1984). The historical background: Examinations and controls in pre-modern China. *Comparative Education*, **20**, 7-26.

HUBERMAN, M. and MILES, M. (1984). *Innovation up close*. New York, NY: Plenum.

HUGHES, A. (1993). *Backwash and TOEFL 2000*. Unpublished manuscript. Reading: University of Reading.

KELLAGHAN, T. and GREANEY, V. (1992). *Using examinations to improve education: A study of fourteen African countries*. Washington, DC: The World Bank.

LAI, C. T. (1970). *A scholar in imperial China*. Hong Kong: Kelly & Walsh.

LI, X. J. (1990). How powerful can a language test be? The MET in China. *Journal of Multilingual and Multicultural Development*, **11**, 5, 393-404.

LINN, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement*, **20**, 179-189.

LINN, R. L. (1992). *Educational assessment: Expanded expectations and challenges* (CSE Technical Report 351). University of Colorado at Boulder: CRESST.

LINN, R. L., BAKER, E. L., and DUNBAR, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, **20**, 8, 15-21.

MACINTOSH, H. G. (1986). The prospects for public examinations in England and Wales. In: Nuttall, D. L.(Ed.). *Assessing educational achievement*. London: The Falmer Press, pp. 19-34.

MADAUS, G. F. (1985a). Public policy and the testing profession: You've never had it so good? *Educational Measurement: Issues and Practice*, **4**, 4, 5-11.

MADAUS, G. F. (1985b). Test scores as administrative mechanisms in educational policy. *Phi Delta Kappa*, **66**, 9, 611-17.

MADAUS, G. F. (1988). The influence of testing on the curriculum. In: Tanner, L. N. (Ed.). *Critical issues in curriculum: Eighty-seventh yearbook of the National Society for the Study of Education*. Chicago: University of Chicago Press, pp. 83-121.

MARKEE, N. (1997). *Managing Curricular Innovation*. Cambridge: Cambridge University Press.

MCEWEN, N. (1995a). Educational accountability in Alberta. *Canadian Journal of Education*, **20**, 27-44.

MCEWEN, N. (1995b). Introducing accountability in education in Canada. *Canadian Journal of Education*, **20**, 1-17.

MESSICK, S. (1975). The standard problem: Meaning and values in measurement and evaluation. *American Psychologist, 30*, 955-966.

MESSICK, S. (1989). Validity. In: LINN R. (Ed.) (3rd Ed.) *Educational Measurement* New York, NY: Macmillan, pp. 13-103.

MESSICK, S. (1992, April). *The interplay between evidence and consequences in the validation of performance assessments*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

MESSICK, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, **23**, 13-23.

MESSICK, S. (1996). Validity and washback in language testing. *Language Testing*, **13**, 3, 241-256.

MORROW, K. (1986). The evaluation of tests of communicative performance. In: PORTAL, M. (Ed.). *Innovations in Language Testing: Proceedings of the IUS/NFER Conference*. London: NFER/Nelson, pp. 1-13.

NOBLE, A. J., and SMITH, M. L. (1994a). *Measurement-driven reform: Research on policy, practice, repercussion* (Centre for the Study of Evaluation Technical Report 381). Arizona State University: CSE.

NOBLE, A. J. and SMITH, M. L. (1994b). *Old and new beliefs about measurement-driven reform: 'The more things change, the more they stay the same'* (CSE Technical Report 373). Arizona State University: CSE.

PEARSON, I. (1988). Tests as levers for change. In: Chamberlain, D. and Baumgardner, R. J. (Eds.). *ESP in the classroom: Practice and evaluation*. Great Britain: Modern English Publications, pp. 98-107.

PETRIE, H. G. (1987). Introduction to evaluation and testing. *Educational Policy*, **1**, 175-180.

POPHAM, W. J. (1983). Measurement as an instructional catalyst. In R. B. Ekstrom (Ed.). *New Directions for Testing and Measurement: Measurement, Technology, and Individuality in Education*, **17**. San Francisco, CA: Jossey-Bass, pp. 19-30.

POPHAM, W. J. (1987). The merits of measurement-driven instruction. *Phi Delta Kappa*, **68**, 679-682.

POPHAM, W. J., CRUSE, K. L., RANKIN, S. C., STANDIFER, P. D., and WILLIAMS, P. L. (1985). Measurement-driven instruction: It is on the road. *Phi Delta Kappa*, 628-634.

RESNICK, L. B. and RESNICK, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In: Gifford, B. R. and O'Connor, M. C., *Changing assessments: Alternative views of aptitude, achievement and instruction*. London: Kluwer Academic Publishers, pp. 37-75.

RESNICK, L. B. (1989). Toward the thinking curriculum: An overview. In: Resnick, L. B. and Klopfer, L. E. (Eds.). *Toward the thinking curriculum: Current cognitive research / 1989 ASCD Yearbook*. Reston, VA: Association for Supervision and Curriculum Development, pp.1-18.

RUNTE, R. (1998). The impact of centralised examinations on teacher professionalism. *Canadian Journal of Education*, **23**, 2,166-181.

SHEPARD, L. A. (1990). Inflated test score gains: Is the problem old norms or teaching the test? *Educational Measurement: Issues and Practice*, **9**, 15-22.

SHEPARD, L. A. (1991). Psychometricians' beliefs about learning. *Educational Researcher*, **20**, 6, 2-16.

SHEPARD, L. A. (1992). What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In: GIFFORD B. R. and O'CONNOR M. C. *Changing Assessments: Alternative Views of Aptitude, Achievement and Instruction*. London: Kluwer Academic Publishers, pp. 301-27.

SHEPARD, L. A.(1993). The place of testing reform in educational reform: A reply to Cizek. *Educational Researcher*, **22**, 4, 10-14.

SHOHAMY, E. (1992). Beyond proficiency testing: A diagnostic feedback testing model

for assessing foreign language learning. *The Modern Language Journal*, **76**, 513-21.

SHOHAMY, E. (1993). *The power of test: The impact of language testing on teaching and learning*. Washington, DC: Occasional Papers published by National Foreign Language Centre.

SHOHAMY, E., DONITSA-SCHMIDT, S., and FERMAN, I. (1996). Test Impact revisited: Washback effect over time. *Language Testing*, **13**, 3, 298-317.

SMITH, M. L (1991a). Meanings of test preparation. *American Educational research Journal*, **28**, 521-42.

SMITH, M. L (1991b). Put to the test: The effects of external testing on teachers. *Educational Researcher*, **20**, 5, 8-11.

SMITH, M. L., EDELSKY, C., DRAPER, K., ROTTENBERG, C. and CHERLAND, M. (1990). *The role of testing in elementary schools* (CSE Technical Report 321). University of California: CSE.

SOMERSET, H. C. A. (1983) *Examination reform: The Kenya experience*. Unpublished report prepared for the World Bank. Revised and published (1987) as *Examination Reform in Kenya*. World Bank discussion paper. Education and Training series.

SPOLSKY, B. (1994). The examination of classroom backwash cycle: Some historical cases. In NUNAN, D., BERRY, V. and BERRY, R. (Eds.). *Bringing about change in language education*. Hong Kong: Dept. of Curriculum Studies, University of Hong Kong, pp. 55-66.

SPOLSKY, B. (1995). *Measured Words*. Oxford: Oxford University Press.

STEVENSON, D. K., and RIEWE, U. (1981). Teachers' attitudes towards language tests and testing. In: CULHANE T. and others (Eds.). *Practice and problems in language testing. Occasional Papers*, **26**. Essex: University of Essex, pp. 146-155.

STIGGINS, R. and FAIRES-CONKIN, N. (1992). *In teachers' hands*. Albany, NY: State

University of New York Press.

SWAIN, M. (1984). Large-scale communicative language testing: A case study. In: SAVIGNON, S. J. and BERNS M. (Eds.). *Initiatives in communicative language teaching*. Reading, MA.: Addison-Wesley, pp. 185-201.

VERNON, P. E. (1956). (2nd Edn.). *The measurement of abilities*. London: University of London Press.

WALL, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language Testing*, **13**, 3, 334-354.

WALL, D. and ALDERSON, J. C. (1993). Examining Washback: The Sri Lankan Impact Study. *Language Testing*, **10**, 1, 41-69.

WALL, D., KALNBERZINA, V., MAZUOLIENE, Z., and TRUUS, K. (1996). The Baltic States Year 12 Examination Project. *Language Testing Update*, **19**,15-27.

WATANABE, Y. (1996). Investigating washback in Japanese EFL classrooms: Problems of methodology. In: WIGGLESWORTH G. AND ELDER C. (Eds.) *The Language Testing Circle: From Inception to Washback. Australian Review of Applied Linguistics Series*, 13. Applied Linguistics Association of Australia: Australia, pp.208-239.

WESDORP, H. (1982). Backwash effects of Language -Testing in primary and secondary education. *Journal of Applied Language Study*, **1**,1, 40-50.

WIDEN, M. F. , O'SHEA, T. and PYE, I. (1997). High-stakes testing and the teaching of science. *Canadian Journal of Education*, **22**, 4, 428-444.

WIGGINS, G. (1989a). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappa*, **70**, 703-13.

WIGGINS, G. (1989b). Teaching to the (authentic) test. *Educational Leadership*, **46**, 7, 41-47.

WIGGINS, G. (1993). Assessment: Authenticity, context, and validity. *Phi Delta Kappa*, **75**, 200-14.

WISEMAN, S. (Ed.). (1961). *Examinations and English education*. Manchester: Manchester University Press.